

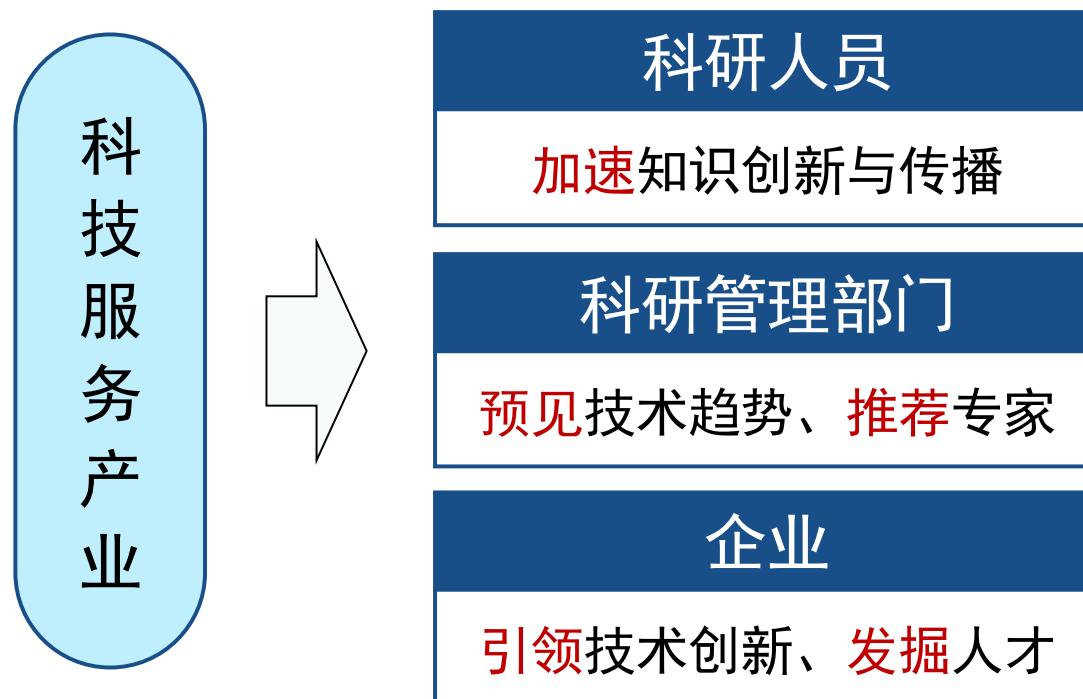


# 知识驱动的科技情报挖掘

Jie Tang  
Tsinghua University

# 科技情报

- 科技情报是从海量科技数据中挖掘出来的**知识创新和传播规律**
- 帮助科研人员、国家科研管理机构和企业提升科技生产力

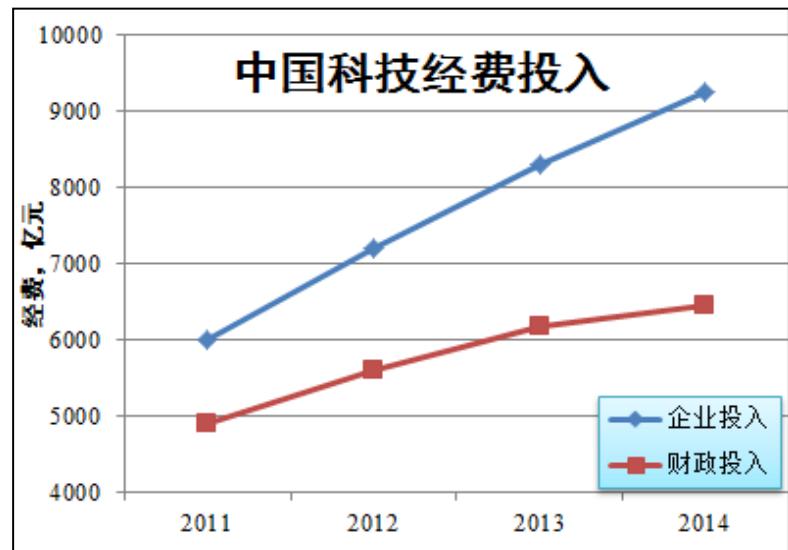


# 情报驱动的科技服务产业

- 谁掌握了科技发展规律，谁就掌握了经济发展引擎！

## □国民经济

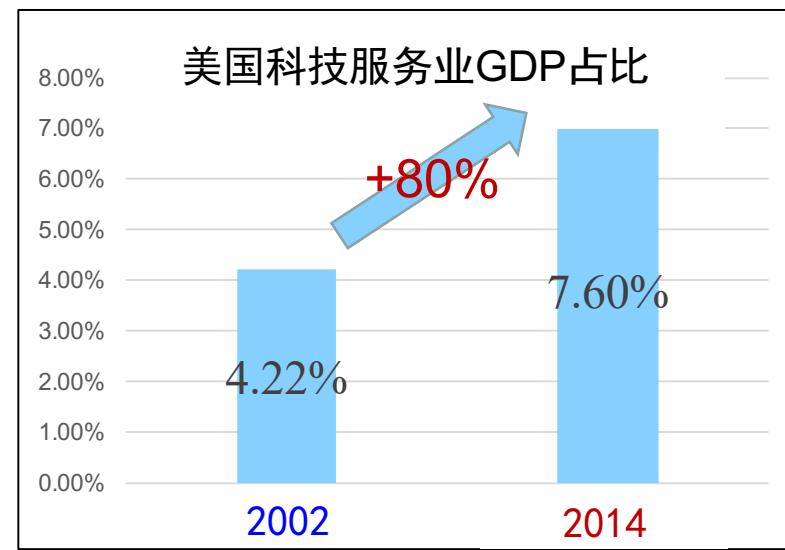
到2020年，科技服务业产业规模有望达到8万亿元



国家统计局, 2014

## □对比美国科技经济

美国科技服务业GDP占比在过去几年增长快速



求是网, 2015

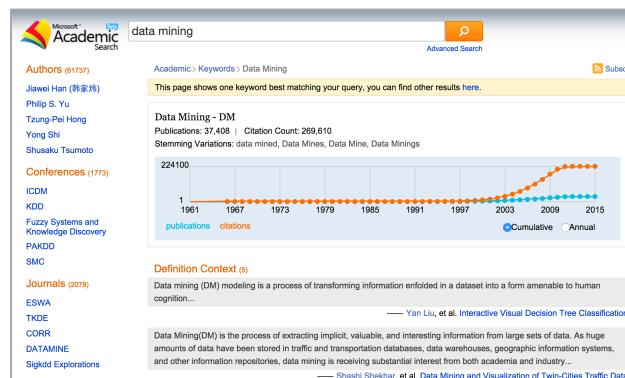
# Background

公司	系统名称	发布时间	特性描述
谷歌	Google Scholar	2004	大规模学术论文索引与快速检索
微软	Academic Search	2014	学术搜索
汤森路透	Web of Science	2008	科学引用指标、交叉学科引用指标
爱思唯尔	Scopus	2004	全学科、交叉学科文献收集

Google Scholar search results for 'data mining'. The results page shows approximately 2,640,000 results. Key snippets include:

- Advances in knowledge discovery and data mining** by M. Fayyad, O. Maimon-Shapiro, P. Smyth... - 1996 - citelike.org
- Data Mining Techniques for marketing, sales, and customer support** by M.J. Berry, G. Linoff - 1997 - dl.acm.org
- The WEKA data mining software: an update** by M.Hall, E.Frank, G.Holmes, B.Pfahringer - ACM SIGKDD ... - 2009 - dl.acm.org

Filters applied: Sort by relevance, Include citations.



Semantic Scholar search results for 'data mining'. The results page highlights 'Data Mining: Concepts and Techniques' by Jiawei Han, Micheline Kamber - MK - 2000. It includes a citation velocity chart and a list of authors. Key sections include:

- Introduction to Data Mining** by Pang-Ning Tan, Michael Steinbach, Vipin Kumar - AW - 2005
- The WEKA data mining software: an update** by Mark A. Hall, Elie Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten - SIGKDD - 2009

# Background

传统情报分析

智能型情报挖掘

构建  
知识图谱

以文献为中  
心的关键词  
情报分析

语义  
挖掘

隐含  
关联

形成知识驱动  
的情报挖掘技  
术体系

挖掘  
人才网络

难以满足科技服务产  
业快速发展的需求

# 科技情报挖掘面临的新挑战

1

情报信息以碎片化的形式  
分散在异构、多源数据中，  
知识获取难度大；

2

情报网络结构复杂、交互  
行为动态多样，挖掘情报  
隐含关联关系十分困难；

3

情报网络规模大、数据多  
维，匹配效率低是智能型  
情报的计算瓶颈。



美国NSF和Army(仅2015年就超过8  
千万美元) - Science of Science



欧盟第8框架H2020(6项关于科技情  
报分析, 共21项学科领域项目)

Science nature

今年截止目前为止的Nature和  
Science杂志中已有23篇关于科技大  
数据挖掘的论文

多个顶尖学术研究机构(MIT, 哈佛,  
康奈尔, 芝加哥等)建立相关团队

# AMiner.cn：知识驱动的科技情报挖掘平台

知识获取  
科技知识图谱构建

关联  
挖掘

隐含关联  
知识到情报挖掘

快速  
算法

语义匹配  
情报到智能服务



# AMiner.cn

MINING DEEP KNOWLEDGE FROM SCIENTIFIC NETWORKS

[Advanced](#)

## Hot Topics

### Data Mining



### Social Network



### Machine Learning



### Deep Learning



### Computer Vision



### Database



136,722,706

RESEARCHERS

101,390,721

PUBLICATIONS

7,854,301

CONCEPTS

133,196,029

CITATIONS

# Open Services —语义搜索

机器学习  
专家

The screenshot shows the AMiner search interface with the query "machine learning". Key features highlighted include:

- Knowledge Graph**: A box highlighting the search results page.
- Demographics**: A box highlighting demographic information: gender (Male 953, Female 47), language (English 272, Chinese 176, Greek 44, French 37, German 28, Korean 12, Indian 8, Japanese 7, Italian 1), and location (USA 263, China 71, United Kingdom 35, Singapore 33, Canada 2, Japan 10).
- Ranking Metrics**: A box highlighting metrics: Relevance, h-index, A-Index, Activity, Diversity, Rising Star, #Citation, and #Paper.
- Rich Semantics**: A box highlighting the semantic details of profiles.
- Machine Learning**: A box highlighting a chart showing popularity over time from 1997 to 2015.

Profile cards shown include:

- Thomas G. Dietterich: h-index 74, #Paper: 376, #Citation: 37790. Similar profiles: 4.
- Guang-Bin Huang: h-index 48, #Paper: 133, #Citation: 25398. School: School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore. Similar profiles: 3.
- Bernhard Schölkopf: h-index 124, #Paper: 639, #Citation: 139285. Director at Max Planck Institute for Intelligent Systems. Similar profiles: 3.
- Pat Langley: Similar profiles: 1.

<https://aminer.cn>

# Open Report – 技术发展报告

[https://www.aminer.cn/research\\_report/articlelist](https://www.aminer.cn/research_report/articlelist)



2018

## 自动驾驶 与人工智能

AMiner 研究报告第七期



科技大学大数据情报中心

2018

## 区块链 基础理论与 应用研究

AMiner 研究报告第三期



2018

## 3D 打印 研究报告

AMiner 研究报告第十一期



2018

## 行为经济学 与人工智能研 究报告

AMiner 研究报告第四期



2018

## 智能机器人 研究报告

AMiner 研究报告第十二期



2018

## 自然语言处理 研究报告

AMiner 研究报告第八期



2018

## 通信 与人工智能研究

AMiner 研究报告第六期



2018

## 机器翻译 与人工智能研 究报告

AMiner 研究报告第五期



2018

## 超级计算机 研究报告

AMiner 研究报告第十期



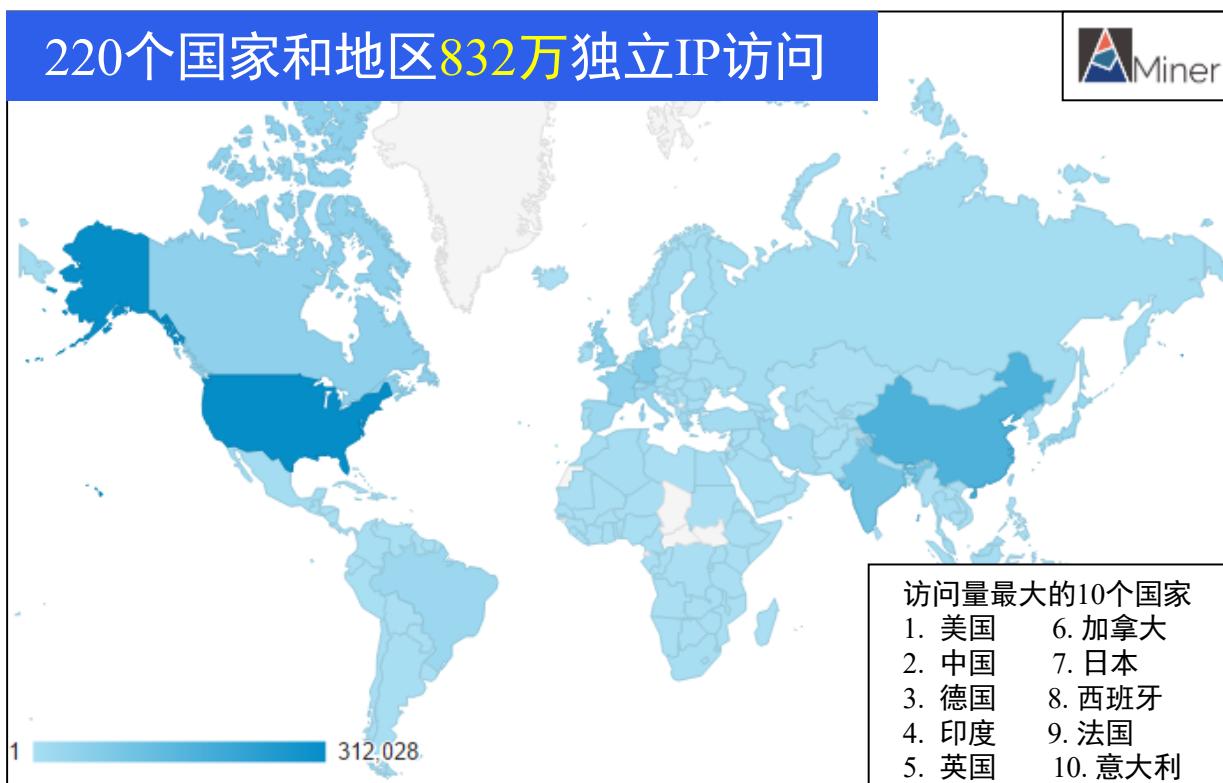
2018

## 人脸识别 研究报告

AMiner 研究报告第十三期

# AMiner访问量及用户分布

## Google Analytics



爱尔兰研究机构  
DERI资深研究员

P. Buitelaar



Exploring Your Research: Sprinkling  
some Saffron on Semantic Web Dog Food

Fergal Monaghan, Georgeta Bordea, Krystian Samp, and Paul Buitelaar

fron cannot be positioned relative to all of these here given the limited space, we now compare Saffron with the most related work: ArnetMiner<sup>8</sup> [6], a well-known state of the art “academic researcher social network search” tool.

ArnetMiner has an emphasis on classification and consists of two main parts. In the first part, probabilistic topic models such as Latent Dirichlet Allocation (LDA) [1] are extended and a unified topic model for papers, authors and conferences is proposed. It seems only the content of the papers is analysed, and structured data, such as social connections, is not considered. They cluster all the words into 100 topics, which is a rather small number considering there are

AMiner is a well-known state  
of the art tool.....

在线运行超过十年

科研数据下载230万次，年均数据访问量超过1100万次



Edward Feigenbaum  
专家系统之父  
图灵奖

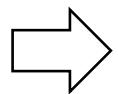
# 核心技术

—How to populate a **semantic**-based profile database for researchers?

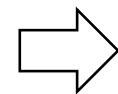


Tim Berners Lee  
WWW创始人  
图灵奖

大数据



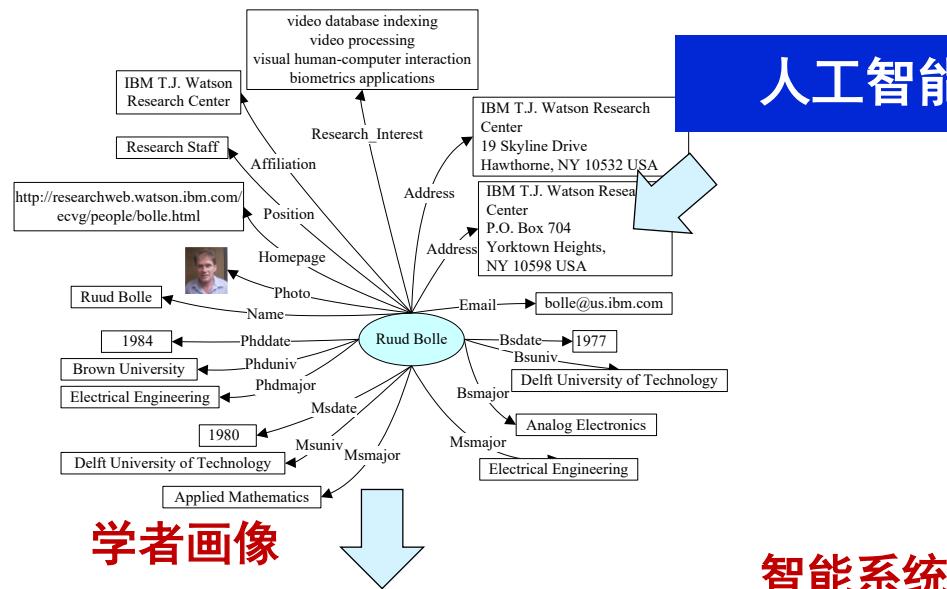
知识



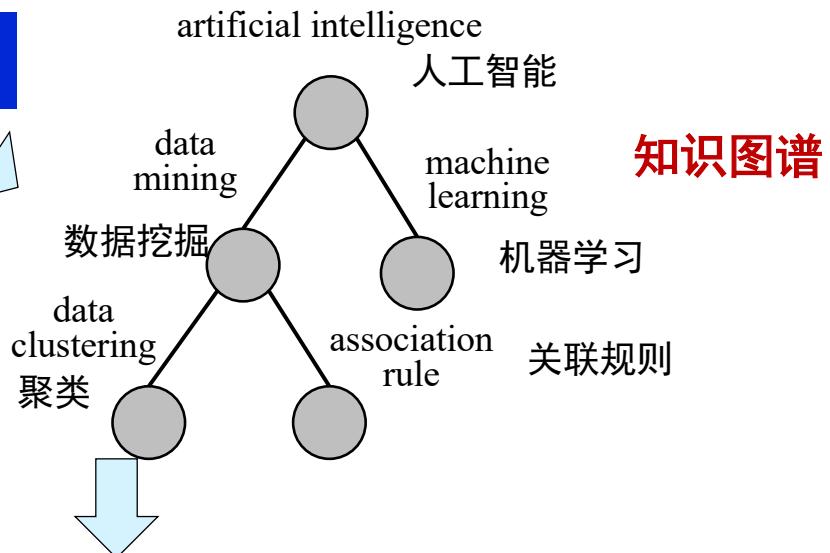
智能

\*人工智能的两个重要阶段：**大规模知识库 + 智能服务**

# Architecture—以AI为例

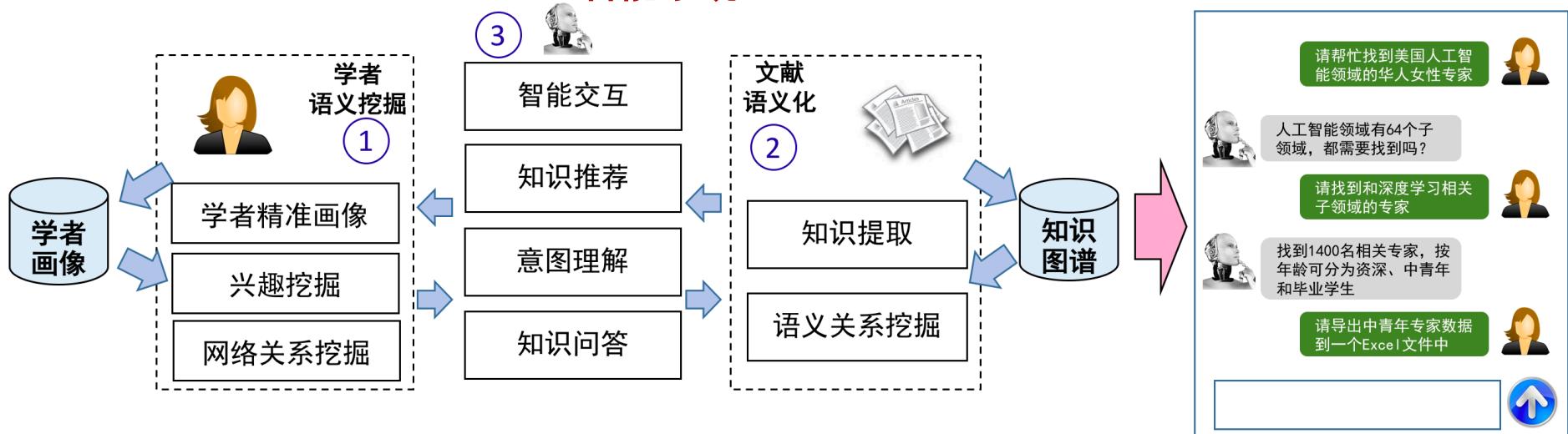


人工智能

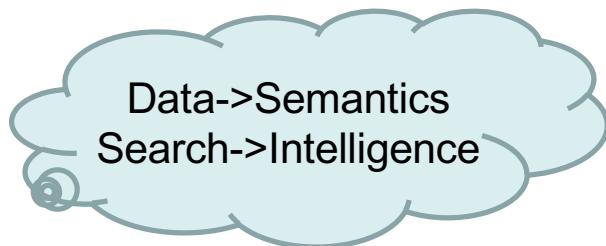


学者画像

智能系统

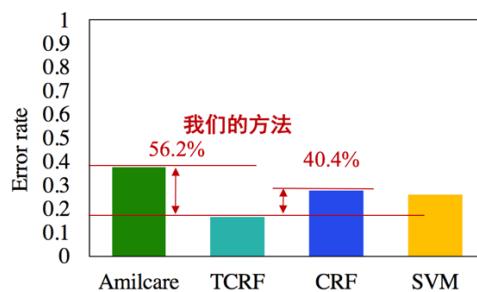


# Technology Overview



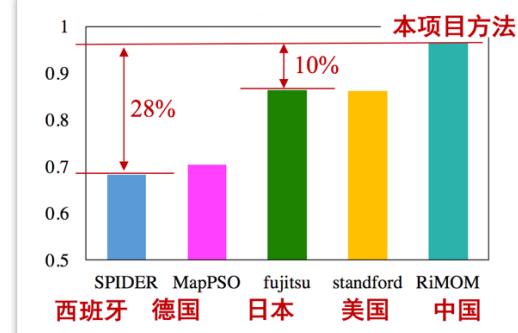
数据语义化

Error rate reduced  
40-56%



语义集成

15 champions in  
the past 7 years



智能服务

Recommendation  
accuracy +161%



Reported by UN

UNITED NATIONS GLOBAL PULSE  
Harnessing big data for development and humanitarian action

**PULSE LAB DIARIES**  
Research Bites: "Inferring User Demographics and Social Strategies in Mobile Social Networks"  
Aug 24, 2014

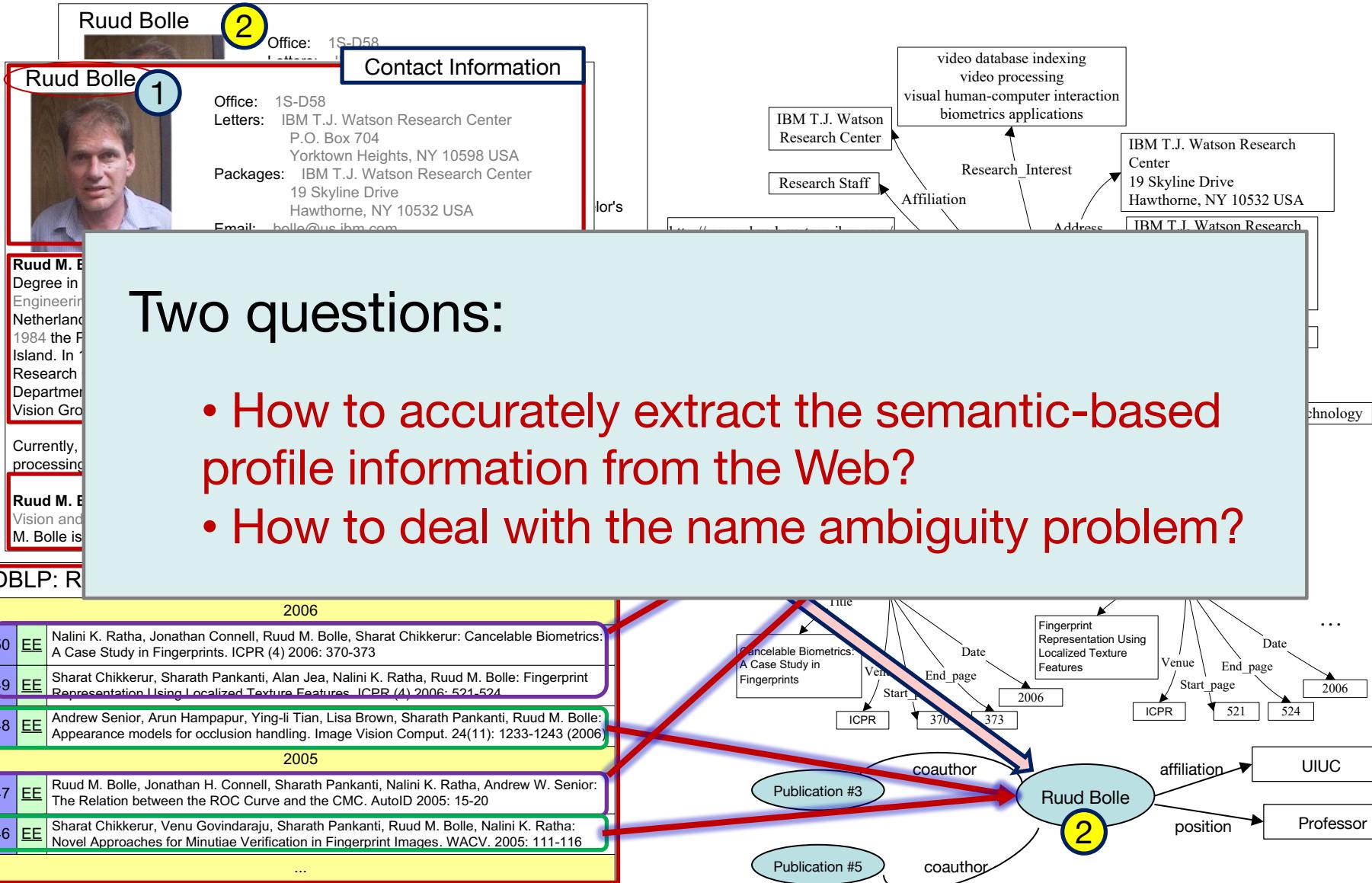
GLOBAL PULSE

ABOUT  
PROJECTS  
LABS  
NEWS  
CHALLENGES  
PRIVACY

#citation>10,000, published ~200 papers on major journals and conferences

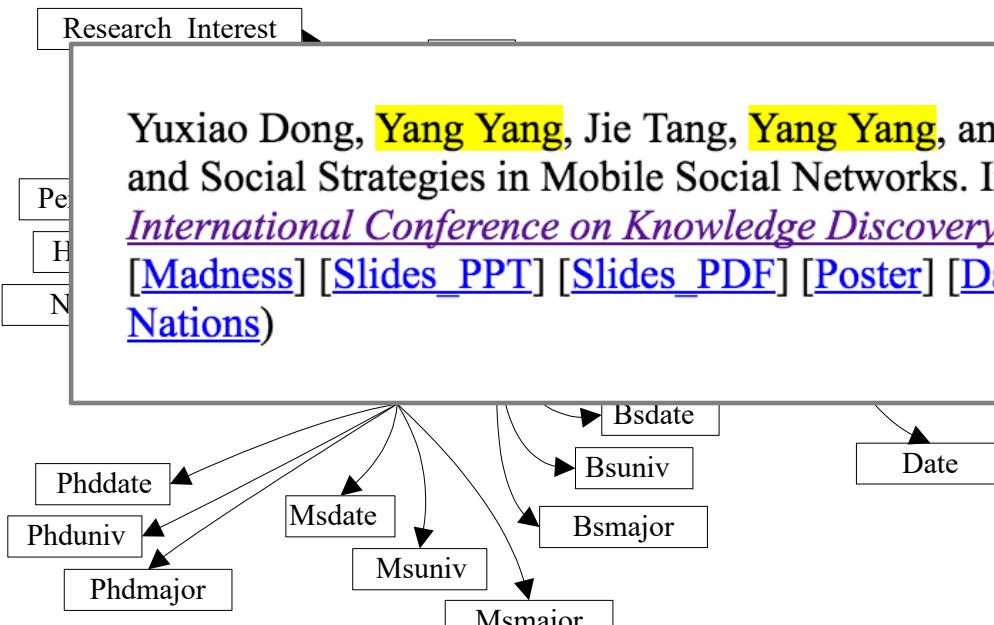
# Extracting Profile Semantics from the Web

(ACM TKDD, WWW'12, ISWC'06, ICDM'07, ACL'07)



# Researcher Profile Extraction<sup>[1,2]</sup>

70.60% of the researchers have at least one homepage or an introducing page



There are a large number of person names having the ambiguity problem

Even 3 “**Yi Li**” graduated the author’s lab

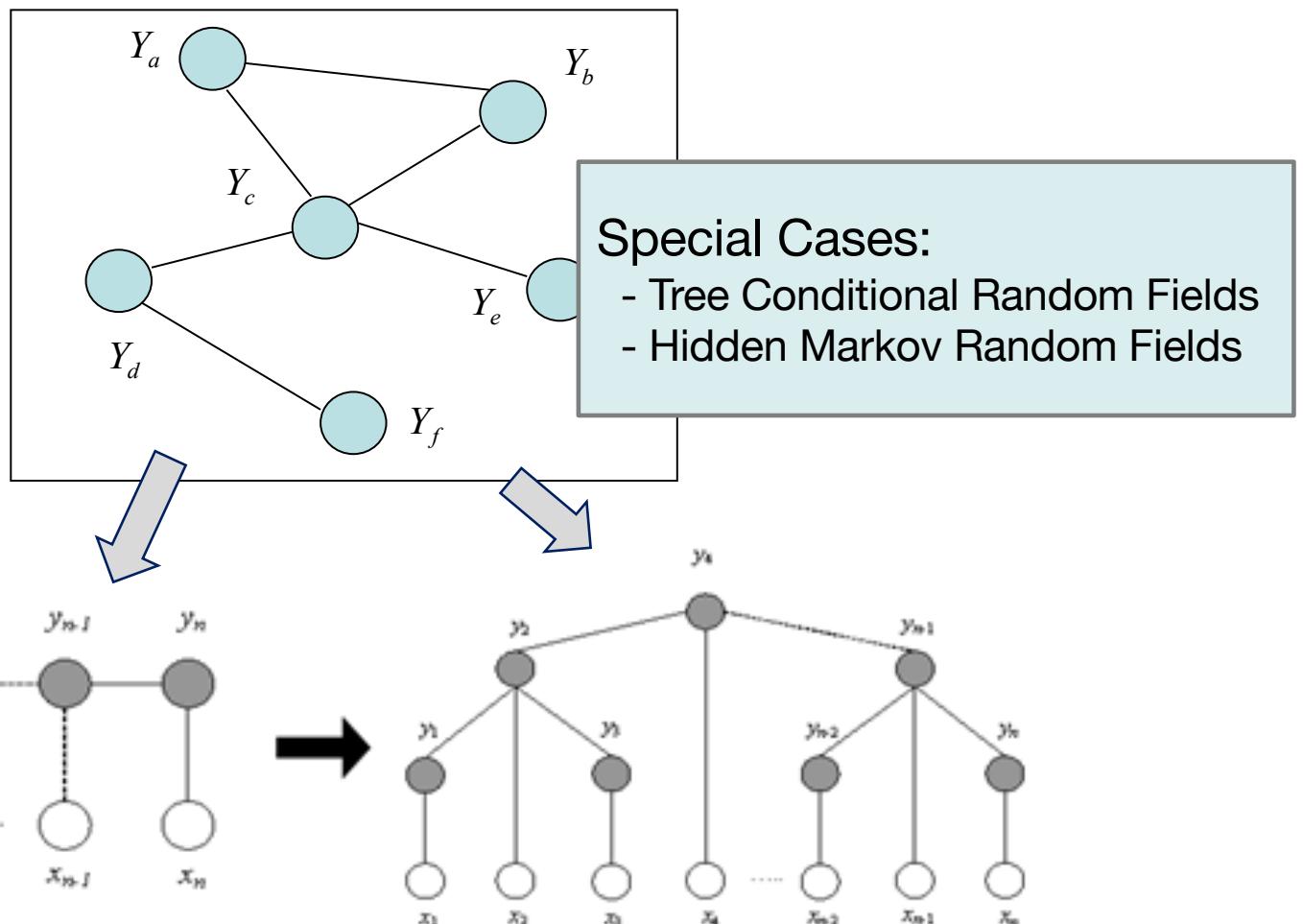
70% moved at least one time

[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. KDD’08. pp.990-998.

# Our Approach Picture

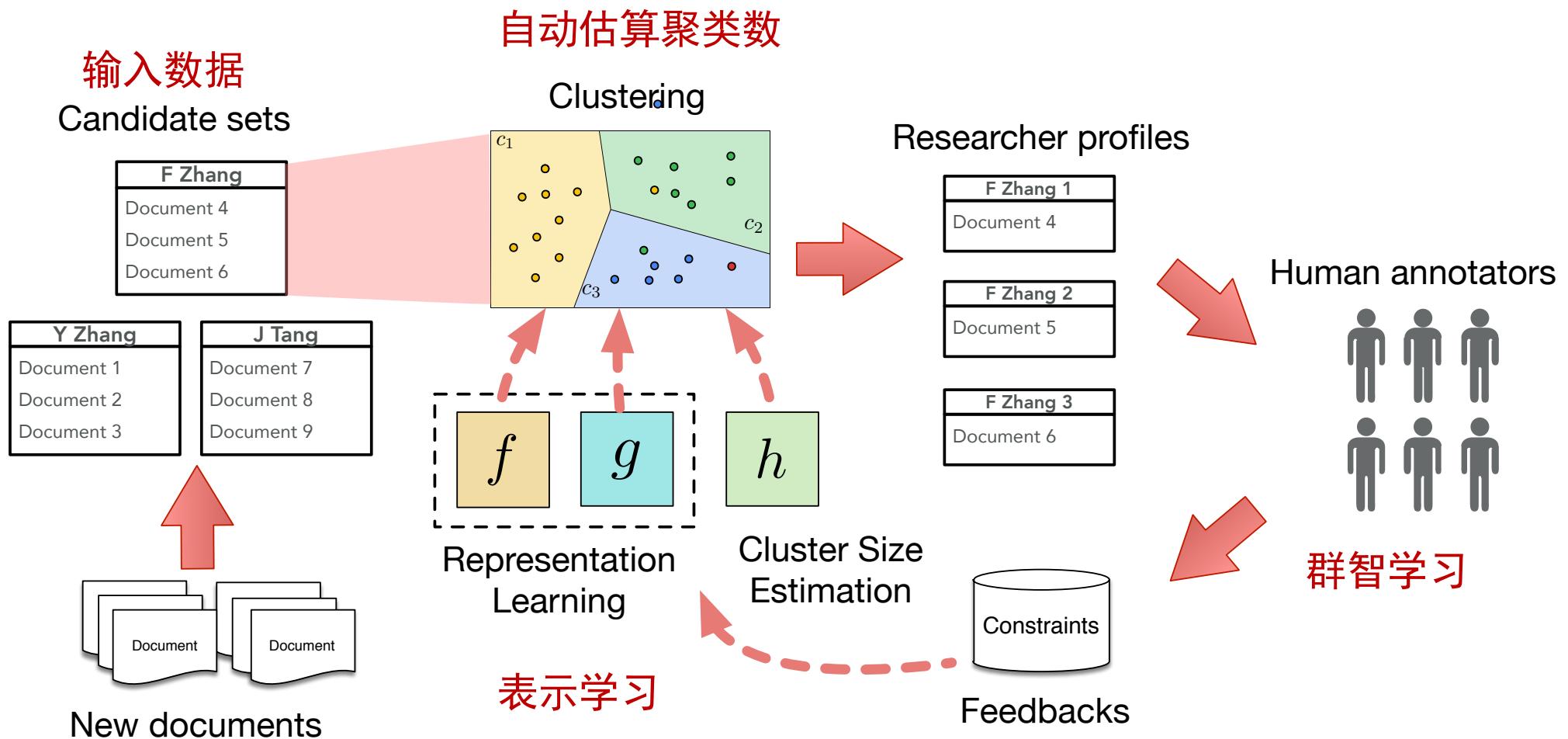
Markov Property:

$$P(Y_i | \{Y_j | Y_j \neq Y_i\}) \\ = P(Y_i | \{Y_j | Y_j \sim Y_i\})$$



$$p(y | x) = \frac{1}{Z(x)} \exp \left( \sum_{e \in \{E^{PC}, E^{CP}, E^{ZX}\}, j} \lambda_j t_j(e, y|_e, x) + \sum_{v \in V, k} \mu_k s_k(v, y|_v, x) \right)$$

# 基于深度学习的在线数据集成框架



[1] Y. Zhang, F. Zhang, P. Yao, and J. Tang. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop. KDD'18. pp. 1002-1011.

# Researcher Profile Database

**ArnetMiner** Home Conference Collaborator Geo Search Topics Download Admin More Welcome jetang Account

Search Experts Search

**Jiawei Han**

Position: Professor  
Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign  
Address: 201 N Goodwin Avenue, Urbana, IL 61801, USA.  
Phone: (217) 332-6903  
Fax: (217) 265-6494  
Email: han@cs.illinois.edu  
Links: [Home](#) [ORCID](#)

**STATISTIC** H-index: 96 Uptrend: 30.46 Diversity: 0.71  
#Papers: 553 Activity: 32.04 Sociability: 726.64  
#Citations: 55885 Longevity: 26 [More Statistics...](#)

**Bio**  
Jiawei Han is computer scientist who specializes in research on Data Mining. He was the 2009 winner of the McDowell Award, the highest technical award made by IEEE. He is currently a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Previously he was a professor in the School of Computing Science at Simon Fraser University. He is an ACM fellow and an IEEE fellow.

**Research Interest**  
Efficient Mining, Spatial Data Mining, Frequent Pattern Mining [Edit Bio](#)

**Education**  
PhD University [Edit Bio](#)

**ArnetMiner** Home Conference Collaborator

Search Experts Search

**Scott**

Position:   
Affiliation:   
Address:   
Phone:   
Links: [Home](#) [ORCID](#)

**STATISTIC** H-index: 96 Uptrend: -4.04 Diversity: 0.22  
#Papers: 195 Activity: 4.86 Sociability: 407.19  
#Citations: 57908 Longevity: 25 [More Statistics...](#)

[Edit](#) [Follow](#)

**Expertise:**  
Wireless network / End-to-end Routing Behavior (80)  
ATM Networks (21)

**M. I. Jordan**

FOAF Follow ALIAS: Michael I. Jordan, Michael Jordan, Michael Irwin Jordan

Position: Professor  
Affiliation: Department of EECS Department of Statistics University of California, Berkeley  
Address: University of California, Berkeley EECS Department 731 Soda Hall #1776 Berkeley, CA 94720-1776  
Phone: (510) 642-3806  
Fax: (510) 642-5775  
Email: jordan@stat.berkeley.edu  
Links: [Home](#) [ORCID](#)

**STATISTIC** H-index: 75 Uptrend: 7.2 Diversity: 0.03  
#Papers: 242 Activity: 11.12 Sociability: 331.69  
#Citations: 44312 Longevity: 23 [More Statistics...](#)

**H. Garcia**

FOAF Follow ALIAS: H. Garcia Molina, H. Garcia-Molina, Hector Garcia-Molina, Hector Garcia Molina

Position: Professor  
Affiliation: Departments of Computer Science and Electrical Engineering.

**See Others:**  
Andreas Peepcke (32)  
H-index: 37  
Jennifer Widom (24)  
H-index: 79  
Barbara Webb (26)  
H-index: 0

**Expertise:**  
Data (115)  
Database Systems (60)  
Database Systems / Automated Data Test (30)  
Mining (23)  
Robot / Hybrid Control (22)  
Library / Information Access

**Conference:**  
VLDB (28) VLDB (27)  
SIGMOD Conference (25)  
ICDECS (23)  
IEEE Data Eng. Bull. (13)  
PVLDB (11)

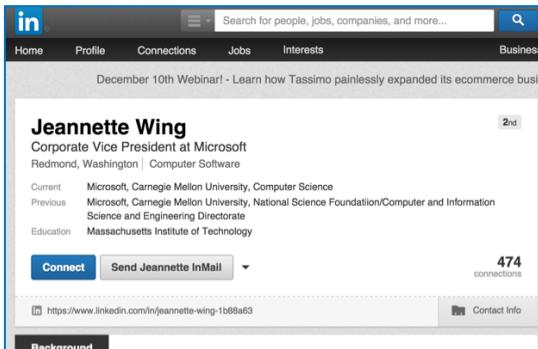
**Research Interest**  
Database Systems, Data Management, Data Warehousing [Edit Bio](#)

[1] J. Tang, L. Yao, D. Zhang, and J. Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages.

# Linking Semantics across Networks

- Identifying users from multiple heterogeneous networks and integrating semantics from the different networks together.

LinkedIn



A screenshot of Jeannette Wing's LinkedIn profile. It shows her title as Corporate Vice President at Microsoft, her location in Redmond, Washington, and her education from Carnegie Mellon University and Massachusetts Institute of Technology. She has 474 connections. A yellow oval highlights the "Same Person" link between this profile and the Wikipedia profile.

WikiPedia



A screenshot of Jeannette Wing's Wikipedia page. It provides a brief biography, her education (S.B. and S.M. in Electrical Engineering and Computer Science at MIT), and her work as a professor at Carnegie Mellon University. A yellow arrow points from the "Same Person" link on the LinkedIn profile to this Wikipedia page.

Same Person

Google Scholar



A screenshot of Jeannette Wing's Google Scholar profile. It displays her citation indices (40, 10, 34, 37) and a bar chart showing citations to her articles over time. A green oval highlights the "Same Person" link between this profile and the AMiner profile.

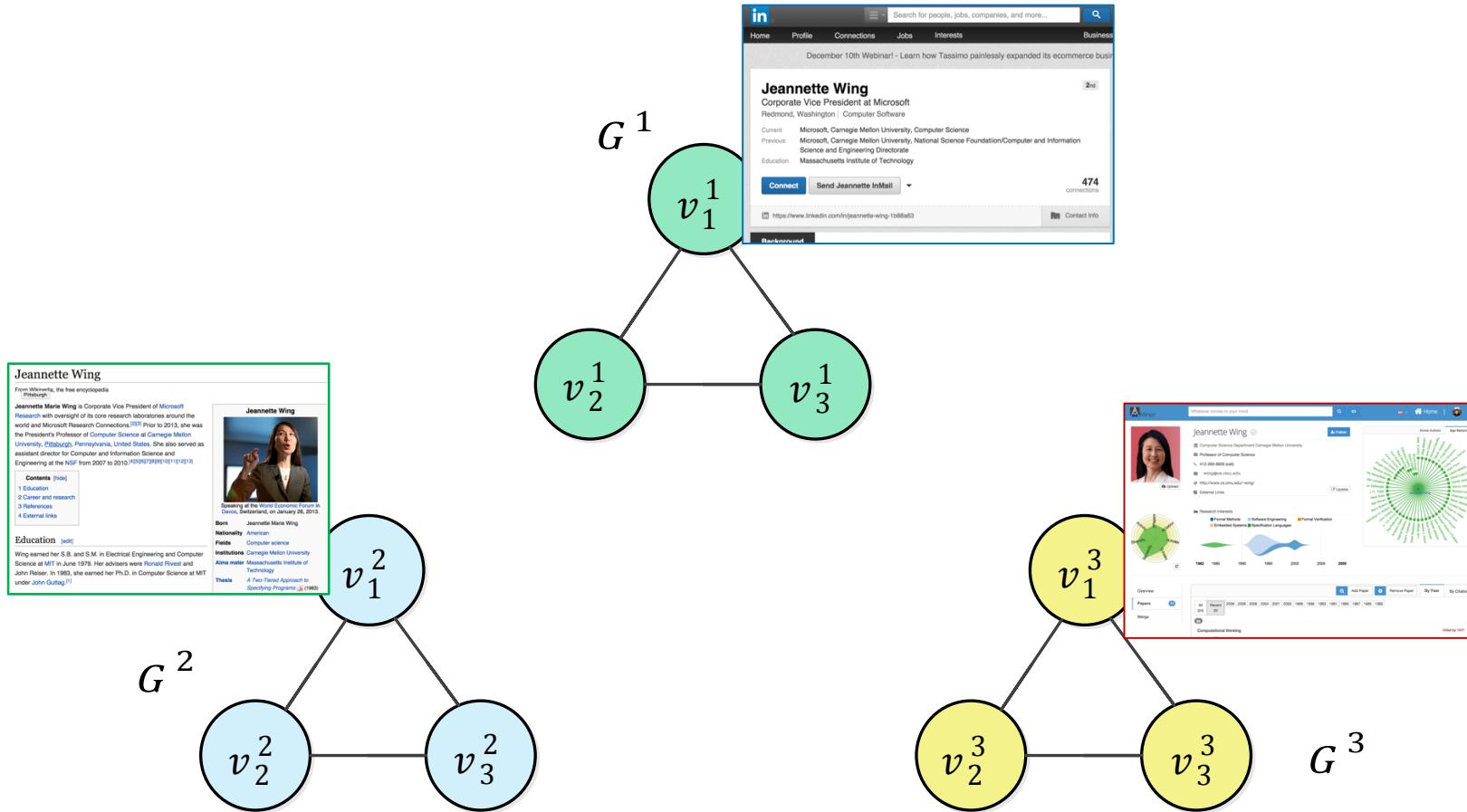
AMiner



A screenshot of Jeannette Wing's AMiner profile. It shows a network graph of her research interests, including Formal Methods, Software Engineering, Embedded Systems, and Specification Languages. A red box highlights the "Same Person" link between this profile and the LinkedIn profile.

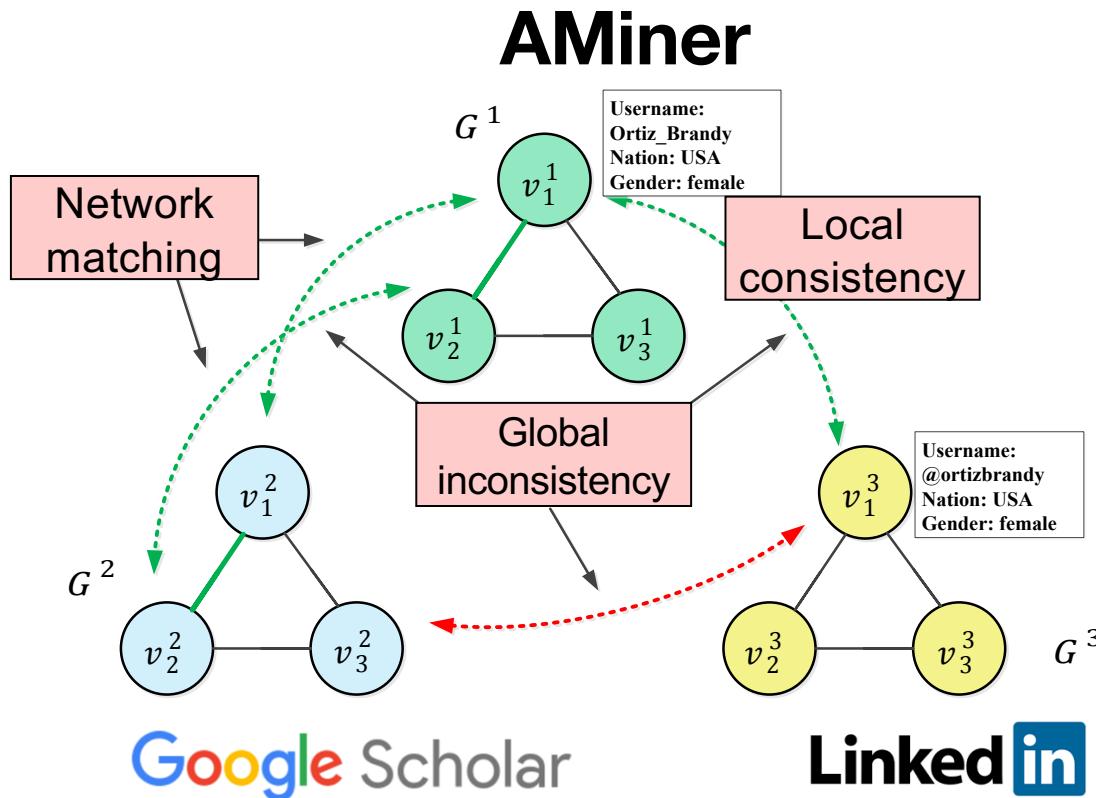
Google Scholar

# Considering the networks...



# Local vs. Global consistency

- Global consistency: matching users by avoiding global inconsistency



**DEFINITION 2 (GLOBAL INCONSISTENCY).** Given a set of social networks  $\mathbf{G}$ , a set of user pairs  $X$  and the corresponding labels  $Y$ , if there exists a sequence of user pairs  $\langle \mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_n} \rangle$ , such that

$$\forall i = i_1, i_2, \dots, i_n, y_i = 1$$

and

$$\forall k = 1, 2, \dots, n-1, \mathcal{V}_{i_k}^2 = \mathcal{V}_{i_{k+1}}^1$$

and

For the pair  $\langle \mathcal{V}_{i_n}^2, \mathcal{V}_{i_1}^1 \rangle$ , if the corresponding label  $y_j = 0$

then we say that the assigned labels  $Y$  causes global inconsistency given  $\mathbf{G}$  and  $X$ .

Avoid “global inconsistency”

# COSNET: Connecting Social Networks with Local and Global Consistency

- **Input:**  $\mathbf{G}=\{G^1, G^2, \dots, G^m\}$ , with  $G^k=(V^k, E^k, R^k)$
- **Formalization:**  $\mathbf{X}=\{x_i\}$ , all possible pairwise matchings and each corresponds to  $y_i \in \{1,0\}$
- **COSNET:** an energy-based model
$$Y^* = \arg \max E(Y, X)$$